Initial Investigation into the Psychoacoustic Properties of Small Unmanned Aerial System Noise

Andrew Christian* and Randolph Cabell†

NASA Langley Research Center, Hampton, VA 23681, U.S.A.

For the past several years, researchers at NASA Langley have been engaged in a series of projects to study the degree to which existing facilities and capabilities, originally created for work on full-scale aircraft, are extensible to smaller scales — those of the small unmanned aerial systems (sUAS, also UAVs and, colloquially, 'drones') that have been showing up in the nation's airspace of late. This paper follows an effort that has led to an initial human–subject psychoacoustic test regarding the annoyance generated by sUAS noise. This effort spans three phases: 1. The collection of the sounds through field recordings. 2. The formulation and execution of a psychoacoustic test using those recordings. 3. The initial analysis of the data from that test. The data suggests a lack of parity between the noise of the recorded sUAS and that of a set of road vehicles that were also recorded and included in the test, as measured by a set of contemporary noise metrics. Future work, including the possibility of further human subject testing, is discussed in light of this suggestion.

I. Introduction

THE ongoing proliferation of small unmanned aerial syatems (sUAS, read "small U.A.S.") has captured the imaginations of many, from single hobbyists to entrepreneurs working for some of the largest companies on the planet. In the United States, a large number of applications for sUAS are now open for commercial exploration given the FAA's changes in policy over the past several years.[1] It may be that, in a short while, communities across the US will be inundated with new classes of noise due to sUAS operations that they had not before encountered.

The Design Environment for Novel Vertical Lift Vehicles (DELIVER) project at NASA has been working to determine the feasibility of producing a conceptual design tool for sUAS that encapsulates NASA's capabilities in the field of full-scale rotorcraft and fixed-wing aircraft design. This tool would bring together estimations of the performance of a proposed vehicle (speed, range, etc.) as well as the environmental impact (i.e., noise and annoyance; though also efficiency, emissions, etc.).

In order to incorporate annoyance into such a tool, some relationship between the predicted physical sounds generated by sUAS must be related to the annoyance that the sound would be expected to generate in human listeners. To date, there have not been any objective studies published to gain even a coarse view of annoyance due to sUAS noise specifically. Further, it is clear that the noise of these machines does not resemble, qualitatively, the noise of contemporary aircraft. This difference in sound quality introduces an unknown factor into the prediction of the resultant annoyance. This paper describes a line of research which seeks to remedy that shortcoming.

Research Premise

In order to formulate a psychoacoustic test, it is necessary to first define a plausible research question that such a test might answer. A common early expectation was that multi-rotor sUAS would be used to deliver packages to residential communities (see, e.g., Clarkson's discussion of the topic from late 2015[2]). A reasonable expectation might be that noise from such sUAS operations will be able to be directly compared

*Aerospace Technologist, Structural Acoustics Branch, M/S 463, AIAA Member.
†Branch Head, Structural Acoustics Branch, M/S 463.

to the noise produced by the contemporary machines used to perform the same task in society. That is, the assumption can be tested that there is nothing about the sound of sUAS that implies that it need be treated in some special way *vis a vis* noise from, for instance, a delivery truck in a residential neighborhood. This premise, while quite broad, is easily tested through "single-event" psychoacoustic methods that have been in use for decades (see, e.g., Shepherd[3]).

Generally, these methods involve collecting/creating various samples of noise to be tested. These sample sounds are then used to formulate a test to which a cross–section of the public will be invited as human subjects. Those subjects will be exposed to the noise samples one by one, and their response, in terms of annoyance, will be solicited to each noise individually. The resultant data is then analyzed in various ways in order to answer targeted research questions.

Organization

This paper consists of three main parts corresponding to the steps that have been outlined: It first describes an effort to record the noise of various sUAS in operation. Second, it describes the initial psychoacoustic test on human subjects that employed these recordings, as well as other sounds. Last, it details the initial analysis of the resulting human subject data and offers two interpretations that will help guide future research in this vein.

II. Sound Collection

This section describes the collection of sounds for inclusion in the psychoacoustic test. First, it describes the recording of a range of multi-copter sUAS executing flyover operations above an array of microphones on or near the ground. It then describes the recording of a number of road vehicles that are meant to be representative of private and commercial operations in residential communities. Finally, the sources of several auralizations — completely synthetic sounds based on aeroacoustic predictions of sUAS/aircraft operations — that are included in the test are discussed.

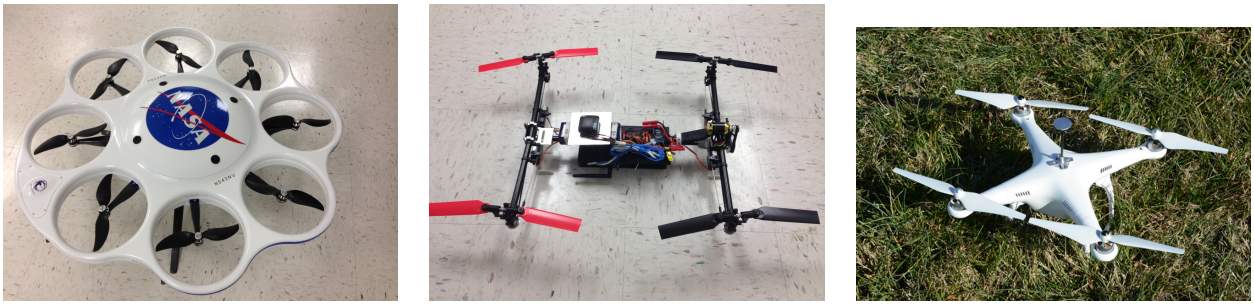
A. Recordings

The bulk of the test was comprised of recorded sounds which were collected at three locations around the U.S. between September 2016 and January 2017.

1. *Oliver Farms II*

The first set of recordings involved NASA–owned, commercially available sUAS. These were recorded during flights that took place on a small grass airstrip, referred to as Oliver Farms, bordered by sorghum fields near Smithfield, VA during late September, 2016.

Photos of the three vehicles flown at Oliver Farms are shown in Fig. 1. They include, from left to right in the figure, the Drone America DaX8 octocopter, the Stingray 500 variable pitch quadcopter (VPV), and the DJI Phantom 2 fixed–pitch quadcopter. Characteristics, including vehicle weight, type, and control method during the flights, are listed in Table 1.



(a) DaX 8 (b) Stingray 500 (VPV) (c) Phantom 2

Figure 1. Photos of sUAS recorded at Oliver Farms.

Table 1. Attributes of sUAS recorded at Oliver Farms.

Vehicle Name	Type	Weight (kg)	Control Method
DaX 8[4]	fixed-pitch octocopter	∼8	autopilot
Stingray 500[5] (VPV)	variable-pitch quadcopter	∼2	manual
DJI Phantom 2[6]	fixed-pitch quadcopter	1.6	autopilot

Nominal flight trajectories for the vehicles consisted of straight–and–level flyovers at 5 and 10 m/s forward flight speed at various altitudes above ground level (AGL). The flight paths were aligned with the long dimension of the runway. If an sUAS was incapable of 10 m/s, the vehicle's highest practical speed was used. Actual speeds and altitudes obtained during the recordings varied depending on vehicle and pilot/autopilot capabilities. Vehicle state information, including roll, pitch, yaw, and GPS position, was recorded with a detachable 3DR Pixhawk flight data acquisition system (FDAS).[7] This system recorded position at a 5 Hz rate with nominal GPS accuracy of 4 m.[8]

The control method in Table 1 refers to whether the vehicle was remotely controlled by a human pilot (manual piloting) or by an on-board flight controller (autopilot). For manual piloting, the pilot stood on the side of the flight path and did his best to maintain a steady altitude and flight speed over a 120 to 300-meter long flight path centered on the runway. In this case the length of the flight path depended on the pilot's comfort with the vehicle's flying characteristics and visibility at the flight path extremes. For example, the Stingray 500 vehicle was difficult to handle and as a result the flyovers for that vehicle were shorter and had greater variability than the other vehicles.

The DJI Phantom 2 was flown with three different bladesets in order to capture possible noise signature variations due to the blades. These blades include the standard OEM blades delivered with the vehicle, a carbon fiber set, and a "slow flyer" propeller manufactured by Advanced Precision Composites. Differences between these bladesets are described in more detail by Zawodny.[9]

The methods for recording acoustic and flight data were similar to those used during previous NASA sUAS recording efforts.[10] The sUAS flyover noise was recorded with three microphones. Two of the microphones were placed directly beneath the flyover path of the sUAS; one on a tripod 1.2 m AGL and the second on a 0.4 m diameter rigid plastic ground board directly below the tripod microphone. The third 'sideline' microphone, also on a rigid ground board, was displaced from the runway centerline 10 m, along the short dimension of the runway. GPS coordinates of the microphones were measured using a u-blox EVK 7-P evaluation kit.[11] All recordings were made using GRAS 40AQ random incidence 1/2" prepolarized condenser microphones coupled with GRAS 26CA constant current power preamplifiers. The preamplifiers were connected to a GRAS 12AX 4-channel power module that provided the constant current needed. For the Oliver Farms test, the microphone responses were digitized using a 5-channel National Instruments NI-4432 USB module at a 20 kHz sample rate. The resulting sampled data were streamed to the hard drive of a connected laptop computer.

The audio recording hardware simultaneously sampled an analog IRIG-B timecode signal to enable time synchronization between the GPS vehicle position data and the acoustic data. The analog timecode signal was demodulated and decoded to obtain a UTC-synchronized time signal for the sampled acoustic data. Seventeen seconds were subtracted from the GPS time in the vehicle state data to account for leap seconds in the GPS data as of late 2016.[12]

Wind conditions were generally calm for these flights, with winds less than 10 knots. A large number of cicadas and birds were present in the woods that bordered the field. This was an unfortunate noise source that necessitated attention during the test design phase, as discussed below.

2. San Diego

Additional sUAS recordings were taken later in 2016, in the Cleveland National Forest, about 35 miles NE of San Diego, CA. These flights were conducted with help from Straight Up Imaging, a San Diego firm which builds and operates sUAS for imaging/surveying purposes. These flights recorded the SUI flagship Endurance sUAS, shown in Figure 2 (referred to simply as "SUI" for the remainder of this document). This model weighs approximately 3.2 kg unloaded. The Endurance performed auto–piloted straight–and–level

flyovers at 5 and 10 m/s speed, and at 20, 30, 50, and 100 m AGL. The SUI flyovers were much more tightly controlled than the flyovers at Oliver Farms. Winds were calm on the day of the test, and ambient noise was significantly lower than at Oliver Farms.

The microphones and associated equipment were equivalent to that used at Oliver Farms, except that the microphone responses were recorded using a Tascam DR-701D 6-track field recorder in uncompressed PCM format.[13]

Figure 2. Photo of the SUI Endurance quadcopter.

3. Langley Research Center

Ground vehicle recordings were made at NASA Langley on a quiet weekend in January. This set of recordings captured drive–bys of four road vehicles, all in good mechanical condition, shown in Figure 3 and described in Table 2, on a long stretch of flat and straight road. The target test condition was a 10 m/s drive–by of the tripod–mounted microphone, which was approximately 10 m from the centerline of the vehicle's path. The drivers were instructed to maintain a constant speed while passing the microphones to minimize engine noise associated with acceleration. The recordings were dominated by tire noise, as well as some low frequency engine noise for the larger vehicles. As it was winter in Virginia, there was considerably less background noise from local fauna than at Oliver Farms. The recording equipment was identical to that used in San Diego. Ground vehicle position information was recorded with the u-blox EVK 7-P evaluation kit.[11]

Table 2. Ground Vehicles

Make	Model	Description
Subaru	Impreza Sport	Passenger hatchback
Ford	Econoline 350	Utility van
International Harvester	MaxxForce DT Dura*Star*	20' box truck (diesel)
Grumman	Kurbmaster/Utilimaster	Step van

B. Auralizations

Although the results discussed in this paper concern the recorded vehicle sounds only, the sounds presented to the test subjects included additional computer–generated sounds, or auralizations. These auralizations were included for comparison with previous human subject tests from which they were taken, and in anticipation of follow–on testing using auralizations of sUAS.

(a) Subaru Impreza

(b) Utility Van and Box Truck

(c) Step Van

Figure 3. Road vehicles included in Langley recordings.

1. Quadcopter

Sounds were included from a group of auralizations that were generated based on a computer simulation of quadcopter flight through atmospheres of varying depths of realism.[14] The dynamics of the simulated quadcopter were modeled so that the various physical forces acting on it could be added/removed easily in order to study the isolated/combined effect they had on the resultant acoustics. These effects include first– and second–order drag forces, the effect of atmospheric turbulence, and the effect that manufacturing tolerances at the component level might have on the operation of a quadcopter.

These auralizations have been used in demonstrations at Langley, so that a large amount of anecdotal response data to the noise has been gathered already. It is therefore instructive to include these sounds in a formal study of psychoacoustic annoyance, though they are not expected to be directly comparable to their real/recorded counterparts. Additionally, it is likely that this auralization capability will find use in future psychoacoustic studies, as it gives the experimenter the ability to control the physical and acoustical properties of UAV flight that may not be possible in the real world.

2. DEP Vehicle

Another group of auralizations that were included in the test came from the Distributed Electric Propulsion psychoacoustic test that took place in 2015.[15] These sounds represent a small civil aviation plane that employs a varying number (6, 12, or 18) of high lift/low noise electric propellers on the leading edge of the wings. These propellers are controlled in such a way that there may be small prescribed differences in rotational speed between adjacent propellers. In addition, the effect of atmospheric turbulence on the phase–stability of the sounds from the various propellers was modeled. The resulting phase and frequency differences between these sound sources can give the predicted noise of the DEP vehicle a variety of unique characteristics that are, perhaps, most appropriately described as 'Jetsons–like.'

The DEP psychoacoustic test employed a similar testing modality to that used here — subjects were played a large number of single sounds and asked to rate their annoyance on a single continuous scale. Accordingly, these sounds were primarily included in order to study the parity between the responses from the DEP test and those found in this sUAS test. This analysis may be the subject of future efforts, and details of the responses to the DEP sounds are not included in the analysis presented here.[16]

III. Psychoacoustic Test

The next section details the formulation and execution of the psychoacoustic test based on the collection of recordings/auralizations discussed above. There are several steps to this process: First, one or more specific research questions to be answered by the test must be formulated. Next, a subset of the recorded sounds designed to address these questions must be determined. Then, the process by which individual sounds are distilled from their 'raw' recorded form into one ready to be presented to subjects is presented. Finally, details of the execution of the test and the facility used are given.

A. Research Questions

The research questions that will guide the selection of the test signals are three–fold:

1. How do subjects interact with the task of rating annoyance in general? At a basic level, this question can be answered by observing, for instance, the amount of variance that exists between answers of the same sound when that sound is repeated at various times during the test. This also includes observation of the variance that exists between subjects for the same sound. The technique of analysis of variance (ANOVA) is used to answer these questions.

2. How do the operational factors impact annoyance? Examples of these factors are the vehicle type, speed, and altitude. Again, ANOVA would be used primarily to answer this question.

3. Is there an observable difference between the annoyance produced by the set of sUAS used in this study and that produced by the road vehicles used? This question is most akin to the premise of this research as discussed in the introduction. It will be addressed here by the application of several forms of linear regression.

Given that subjects will be simply asked to rate their annoyance to a large set of sounds on some subjective scale, there may be many more questions that various types of further analysis may support, though no conjecture as to what those might be are offered here.

B. Signal Selection

Using these research questions as guidance, the set of sounds to be included in the psychoacoustic test was formed from the set of recordings. The final selection is shown in Table 3 on page 7. Every row in Table 3 represents a single flyover sound. The ID numbers used in Table 3 come from those given to the recordings, so that they do not form a contiguous set in the test, nor are they listed in order in the table. Repeated sounds were given unique ID numbers, for instance numbers 1, 51, and 61 all represent the same sound to be presented to the subjects multiple times throughout the test. In this way, the first research question can be addressed by all of the rows in Table 3 that have multiple IDs. Further, rows 1-3 and 3-6 constitute sets of recordings made of the same nominal parameters (vehicle type, speed, etc.), in order to explore whether repeated observations of the same operations produce equal annoyance.

All recorded sounds are taken from the tripod microphone except for IDs that are in the range of 100-199, which are taken from the sideline ground board microphone. The last two digits of these IDs correspond to their counterparts of the 0-99 range (e.g., ID 1 is the same flyover as ID 101). The inclusion of this set is meant to determine whether observations at slightly different locations (and without interference from the ground plane) produce significantly different results. This was only explored for the SUI vehicle.

For the sUAS included in the test, the 'Configuration' column of Table 3 indicates the blade set used. For the SUI vehicle, OEM–2 refers to the two-bladed configuration, and OEM-3 refers to the three-bladed configuration. For the Phantom 2, OEM refers to the standard bladeset, CF to the carbon fiber blades, and APC to the Advanced Precision Composites blades. The height and speed values listed are nominal, as indicated in the previous discussion regarding the differences between manual and autopilot control.

All of the SUI samples were presented to the subjects at their original recorded level. For the other sUAS included in the test, the level was manipulated in order to produce a range of around 15 $dB_{A,Max}$ per–vehicle. As some sUAS were naturally quieter than others for the same flight condition, this produced a test that was primarily contained within a 20 $dB_{A,Max}$ range. This spread was intended to be wide enough to provide sufficient range for the subjects and for the linear regression, but not so wide as to generate contraction effects on the response scale (see, e.g., Bech[17]).

Gains were also applied to the road vehicle recordings. Given that they were all recorded at the same operating condition, there was very little observed variation between the recordings of the same vehicle. The span within and between road vehicle samples was made to be roughly equal to that of the sUAS samples in terms of dB_A. For the road vehicles, the values in the height column indicate the nominal distance from the microphone to the centerline of the vehicle.

Once a recording was chosen for inclusion, a start and stop point were determined to extricate the sample from the larger recording. These points were determined by observing the maximum dB_A level reached by the event, and choosing points in time to either side of that maximum that corresponded to the same dB–down level. This level was set between 10 and 20 dB down in nearly all cases. For example, if the maximum was 65 dB_A, then the sound may have been selected so that it started and stopped around 50 dB_A. This level was often determined by the presence of extraneous sounds (e.g., birds chirping) that became clearly audible as the level of the flyover decreased. In some cases, particularly the high–altitude sUAS cases, points at less than 10 dB down were selected due to the extreme lengths of sounds that would have been produced by following the strategy. This process resulted in the lengths of the samples as noted in Table 3.

Four versions of the quadcopter auralization were included. As indicated in Table 3, each was presented to the subjects three times. The conditions 1-4 indicate various realistic effects included in the flight dynamics simulations that were run as input to the auralizations (again, see Christian[14]). These correspond to, cumulatively: 1. No dynamical effects. 2. Drag effects on the body and rotors. 3. A model of turbulence acting on the sUAS. 4. Sources of random error included between the thrust coefficients of the four rotors.

A set of nine various auralizations of the DEP vehicle were included. These were of simulated flyovers at 300 m AGL, with a speed of 31 m/s. This gives a receiver–angle time history somewhat similar to that of the sUAS recordings.

Table 3. Index of sounds included in the psychoacoustic test. See text for full description. N.B., The ID numbers are not necessarily in ascending order.

ID Number(s)	Vehicle	Configuration	Height (m AGL)	Speed (m/s)	Sound Length	Gain Applied
1, 51, 61	SUI	OEM-2	20	5	25	0
2, 52, 62	SUI	OEM-2	20	5	26	0
4, 54, 64	SUI	OEM-2	20	5	27	0
101, 151, 161	SUI	OEM-2	20	5	25	0
102, 151, 162	SUI	OEM-2	20	5	26	0
104, 154, 164	SUI	OEM-2	20	5	27	0
5	SUI	OEM-2	30	5	34	0
10	SUI	OEM-2	20	10	11	0
13	SUI	OEM-2	30	10	17	0
17	SUI	OEM-2	50	5	29	0
20	SUI	OEM-2	100	5	48	0
24	SUI	OEM-3	20	5	22	0
30	SUI	OEM-3	30	5	22	0
32	SUI	OEM-3	50	5	35	0
35	SUI	OEM-3	100	5	40	0
105	SUI	OEM-3	30	5	34	0
110	SUI	OEM-3	20	10	11	0
113	SUI	OEM-3	30	10	17	0
117	SUI	OEM-3	50	5	29	0
120	SUI	OEM-3	100	5	48	0
124	SUI	OEM-3	20	5	22	0
130	SUI	OEM-3	30	5	22	0
132	SUI	OEM-3	50	5	35	0
135	SUI	OEM-3	100	5	40	0
204	DaX 8	OEM	20	5	18	0
207	DaX 8	OEM	20	5	20	-12
212	DaX 8	OEM	40	5	28	0
213	DaX 8	OEM	40	5	30	-12
220	DaX 8	OEM	55	5	31	0
221	DaX 8	OEM	55	5	35	-12
242	Phantom 2	APC	20	10	12	0
245	Phantom 2	APC	20	5	13	0
246	Phantom 2	APC	5	5	10	0
250	Phantom 2	APC	5	10	10	0
257	Phantom 2	APC	20	10	8	-6
262	Phantom 2	CF	10	10	16	0
267	Phantom 2	CF	20	5	16	0
269	Phantom 2	CF	20	10	14	0
264	Phantom 2	CF	5	10	10	-8
272	Phantom 2	CF	20	10	15	-8
287	Phantom 2	OEM	20	5	17	0
289	Phantom 2	OEM	20	10	15	0
282	Phantom 2	OEM	7	5	9	0
382	Phantom 2	OEM	7	5	9	-8
387	Phantom 2	OEM	20	5	17	-8
296	VPV		10	5	14	0
299	VPV		30	5	18	0
300	VPV		10	5	14	0
306	VPV		10	10	13	0
308	VPV		10	10	10	0
404	Subaru		10	10	12	0
407	Subaru		10	10	18	0
408	Subaru		10	10	18	0
457	Subaru		10	10	18	12
458	Subaru		10	10	18	6
410	Step Van		10	10	18	0
415	Step Van		10	10	18	0
417	Step Van		10	10	14	0
465	Step Van		10	10	18	8
467	Step Van		10	10	14	4
418	Box Truck		10	10	17	0
422	Box Truck		10	10	13	-5
423	Box Truck		10	10	14	-10
472	Box Truck		10	10	13	-15
473	Box Truck		10	10	14	-20
424	Utility Van		10	10	12	0
426	Utility Van		10	10	13	0
431	Utility Van		10	10	11	0
474	Utility Van		10	10	12	4
476	Utility Van		10	10	13	-4
601 - 603	Quadcopter*	1	6	6	19	21
611 - 613	Quadcopter*	2	6	6	19	21
621 - 623	Quadcopter*	3	6	6	19	21
631 - 633	Quadcopter*	4	6	6	19	21
701 - 709	DEP*		300	31	13	14

*Indicates an auralization.

C. Presentation Order

Combined, there were 103 non–unique sounds in the test, resulting in about an hour of test time. Through experience with previous tests, it has been determined that a subject should be given breaks in their listening task intermittently — they should not be asked to listen to annoying sounds for an hour straight. The test was therefore broken up into 4 sessions of about 15 minutes each.

It is also desirable to present the test sounds in different orders for each group (of four) subjects. This is to minimize the effect of a possible sequential contraction bias, for instance, where a quiet sound would always be played after a loud one, leading to the former being judged as unreasonably not–annoying due to the consistent contrast between the two sounds. Similar biases may arise when a sound is played always at the beginning or end of the test. Therefore, each group of subjects should be presented with a systematically unique ordering of the sounds in order to alleviate possible biases.[17]

To accomplish this, first, the 103 samples were partitioned into four blocks of relatively equal length. This was done in such a way that similar sounds were not assigned to the same block (e.g., no repeats ever occurred in an individual block). Then, a Latin Square ordering was used to assign one of the blocks of sounds to one of the sessions (see, e.g., Montgomery[18]). As there were more groups of subjects than blocks, the Latin Square pattern was reversed for groups 5-8, and then repeated for groups 9 and 10. For each test session, the ordering of the approximately 26 sounds within a block was randomized uniquely for each session and each group.[a]

D. Signal Processing

Once the sounds were selected for inclusion in the test, a chain of signal processing techniques was implemented in MatLab to condition the recordings into a form in which they could be presented to the subjects.

1. Upsampling

All of the test samples were required to be at 44.1 kHz sampling frequency for playback in the Langley Exterior Effects Room (EER, discussed further below). While the recordings of the SUI Endurance, road vehicles, and the auralizations were already at this resolution, the recordings from Oliver Farms were at 20 kHz having been recorded through the NI ADC. The process to convert the Oliver Farms' recordings to the standard resolution involved 3 steps: 1. Upsample to 100 kHz using a polyphase filter. 2. Create a time-base at the new sampling frequency. 3. Use a cubic spline interpolation scheme to generate the samples at the desired new time base.

In total, this process preserved the information in the Oliver Farms' signals up to 9 kHz. This performance is satisfactory given that the original recordings only contained information up to 10 kHz, and given the high-frequency filtering step described next.

2. Filtering

While the recording system used at Oliver Farms included an NI ADC that was DC coupled, the other recordings were made with the TASCAM recorder, which did not necessarily have the dynamic range of the former. To ensure that the recorder would not clip, for example, due to wind noise, the internal high–pass filter of the TASCAM was set to 50 Hz. To gain parity between the two recording platforms, as well as to filter out unwanted extraneous noise from the Oliver Farms' recordings, a 2^{nd} order Butterworth high–pass filter was designed and applied to those recordings.

For all of the recorded sUAS, 50 Hz worked out to be less than half of the nominal rotor blade passage frequency, implying that the filtering process would have no significant impact on the components of the recording created by the sUAS. It is likely that the road vehicles produced content below this frequency, but it is unlikely that this content would have bearing on the perceived annoyance of the signal (as will be discussed further below).

In addition to this processing, a 1^{st} order Butterworth low-pass filter was applied to the Oliver Farms recordings to remove high frequency white noise observed in some of those recordings. This filter was designed to begin to be effective at 3 kHz. While this frequency is still in the range of interest for human annoyance,

[a]Although this does not preclude the chance of two sounds occurring in a row for two different subject groups, given that there are at least 25 sounds in a session, the chance of this occurring to the extent that it would create a bias is vanishingly small. Additionally, it is always good practice to have at least one random layer in a test design.[18]

the filter was likely not aggressive enough to significantly impact the content of the recording until, perhaps, an octave above its nominal frequency. In this way, the filter smoothly separated the relevant components of the recordings from those that might extraneously impact the annoyance rating thereof.

3. Windowing

After the start and stop points were determined for the individual sample sounds, 2 seconds were added back to each end, during which a fade was applied. This was done to ensure that there were no audible artifacts at the ends of the playback, and so that that the sounds would not appear to 'jump' out from the background in a startling manner.

A suitable fade-in function was found to be F:

$$F(t) = \left[\frac{t}{2}\right]^{\left(2-1.5\frac{t}{2}\right)} \qquad \text{for } t \in [0,2] \tag{1}$$

which multiplies the pressure-time signal (the time-reverse was used for the fade out).

4. Final Considerations

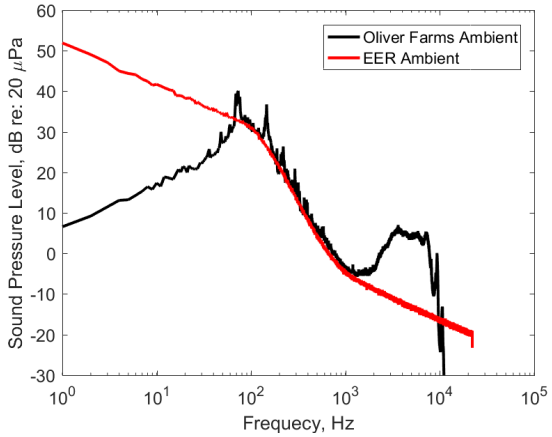
As indicated in the signal selection section, an overall gain was added to some of the signals in order to help the full set span a reasonable range. Once this, as well as the other signal processing tasks, had been applied, the sounds were written to 32–bit floating point wave files (a format that preserved the working units of Pa) for playback in the EER.

E. Test Environment

The test was conducted in the Exterior Effects Room (EER[19]) at the NASA Langley Research Center in Hampton, Virginia. The EER is a small, acoustically–treated auditorium with a 31-channel sound reproduction system capable of simulating 3D spatialization of point-source sounds in real time. The reproduction capability of the EER extends over a compensated frequency range of 20 Hz to 20 kHz and from approximately 23 to 94 dB_A, at which point the playback system is limited to protect the subjects' hearing.

The EER is approved to test 4 subjects at a time as shown in Figure 4(a). The subjects sit in non-consecutive seats located close to the geometric center of the room. Subjects are visually isolated from one another by an acoustically transparent curtain. Between the subjects is a microphone belonging to a sound level meter (SLM) that is both used for calibration as well as to monitor the sound levels during the test.

(a)



(b)

Figure 4. The test facility. (a) NASA employees posing as test subjects in the EER. (b) The background noise condition applied during the test, as compared to ambient noise recorded at Oliver Farms ($f_s = 44.1$ kHz).

1. Ambient Noise Condition

An artificial ambient noise condition was created in the EER based off of the ambient noise levels observed at Oliver farms (the loudest recording site). This ambient was created by filtering white noise into a 3–minute long loop that played through all speakers in the EER. The level of this loop was set to produce an overall sound pressure level of 36 dB_A. The frequency content of this noise is shown in Figure 4(b), as compared to the content of a recording of the ambient noise at Oliver Farms (in which the 'bump' above 2 kHz is due primarily to cicadas).

2. Calibration

Calibration of the sounds for playback in the EER was accomplished by use of the EER SLM. The sample sounds were played back with the EER empty and the ambient noise off. The $dB_{A,Max}$ levels were compared to the maxima predicted by processing the sample sounds in the same manner as the SLM. The difference between this prediction and measurement was minimized.

In this way, calibration factors were obtained separately for the set of flyovers, fly–bys (again, IDs 100-199), and drive–bys. These were separated due to the fact that the difference in the geometry of their playback in the EER created systematic differences of up to 1.5 dB_A between the groups. For the entire set, the standard deviation of the difference between the predicted and calibrated measured sounds was .6 dB.

3. Geometric Processing

The GPS vehicle-position data were used by the EER's spatialization capabilities to create the impression that the sample sounds were traveling along the course taken during their recording. This required the a 'retarded time' correction to be computed — the GPS data produced the location of the source at the instant the noise was being received, not at the instant it was transmitted from that source. Further, all flyovers were normalized so that came directly overhead at their closest point. Finally, a stereoscopic projection (i.e., acoustic fisheye) was applied to the road vehicle sounds after pilot testing in order to prevent them from sounding as if they were "driving through the middle of the room."

F. Test Execution

The test, dubbed WGA–I, took place during the last week of February, 2017. 2 groups of 4 subjects were tested each day for 5 days, resulting in 40 subjects total. 2 subjects did not report for the test on time and were therefore excluded from participation, resulting in a pool of 38 subjects. All subjects listened to all sounds, all responses were successfully captured.

The subjects were recruited from the local community by a contractor — typically from within 50 miles of NASA Langley. The requirements for participation, as provided to the contractors, were to provide subjects that:

- Have no more than 30 dB of hearing loss (relative to reference hearing thresholds in ISO 389-1[20]) over the frequency range of 250 Hz to 4,000 Hz.
- Are within 18 and 50 years of age.
- Create an overall proportion of between 1/3 and 2/3 female participants.

Upon arrival, subjects were given a pre-test hearing screening. Before the test began, subjects listened to a suite of 10 samples selected from the test in order to familiarize them with the breadth of sounds they would be listening to. They then completed 5 practice questions to ensure that they understood the question and how to record their responses using touch–screen tablet computers provided to each subject.

Subjects were asked to rate their annoyance on a single scale shown in Figure 5. This question was formulated based on the recommendation by Fields *et al.*[21] Using this scale, specifically the wording ('Not at all,' etc.), produces subject responses that are linear with the perceptual quantity under study. This facilitates the use of the well–understood linear regression model for data analysis. This scale was stored as a numeric value between 1 and 11, with the even numbers corresponding to the five ticks/words on the scale.[b]

[b]It is good practice to give the subjects the ability to respond past the final tick marks on the ends of the scale.[17]

The test proceeded through the 4 sessions, encompassing all 103 sounds, as described above. Between sessions, subjects were allowed to take an elective break (e.g., to use the restroom). After the test, subjects were required to have their hearing re–screened to ensure that the sounds they had been exposed to during the test did not significantly alter their hearing threshold (a possible indication of noise–induced hearing loss). The total participation time, between pre– and post–test hearing screening was between 1.5 and 2 hours. The test protocol was approved by the NASA Langley Institutional Review Board.

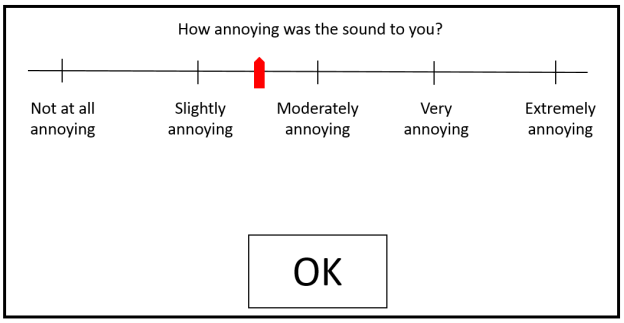


Figure 5. Screenshot of the question posed to the subjects.

Heuristics

Following the test there was a period of time for discussion between the subjects and the researchers. Although this time was not compulsory, the insights given by the subjects, for example, into their decision making processes, can be valuable. For this reason, a selection of observations are given here.

Some subjects reported developing heuristics early on that they then applied to the sounds throughout the rest of the test. There was often disagreement, even contradiction, between the strategies of different people. For example, some subjects reported finding more "high–pitched" sounds more annoying, while others thought "low–pitched" ones were worse, though clearly, these do not necessarily reflect objective measures of the sounds. Some subjects reported the opposite: judging each sound on its own merit, though it is likely that the former was more prevalent.

Another typical comment was that sounds that appeared to linger were judged to be more annoying than those that did not. As long as a sound was not startling, a perception that the sound would "be over with soon" alleviated annoyance. Further, sounds that were described as being "patterned" had a greater tendency to evoke this response than ones that were more qualitatively constant.

No subjects reported having their responses affected by the presentation location of the sound. This was true both when this information was volunteered, and when it was inquired about directly by the researchers. Very few subjects were able to identify the sUAS sounds as coming from 'drones.'

Again, as these responses were not compulsory; they should not be taken as necessarily representative of a significant portion of the subject population.

IV. Initial Analysis

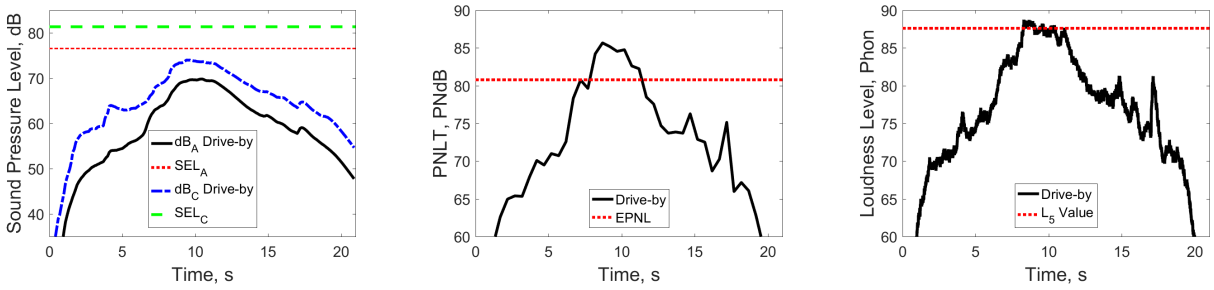
There are several ways to analyze the type of data collected in this test. Only one of these — linear regression — will be discussed at significant length here. Other analyses that will be undertaken in the future, related to the remaining research questions, are discussed briefly at the end of this section.

Linear Regression

Several forms of linear regression were performed between noise metric values computed for the sample sounds and the subject responses to those sounds. Again, the form of the question posed to the subjects was meant to produce responses that vary linearly with annoyance, making more exotic forms of regression likely to be unnecessary.

For this analysis, only recorded sounds were considered; the auralizations, both of the DEP vehicle and the quadcopter, are left out. The inclusion of those sounds in the test was not with the intention of direct comparison between auralized sounds and recordings, therefore any line fit through the data should not attempt to account for the variance between the those classes of sounds. Additionally, repeats of the SUI sounds are omitted; thus of the first 18 sounds listed in Table 3, only IDs 1 and 101 are included in this analysis. This is to ensure that the subject responses are identically distributed between sample sounds, that there is no dependence of one sound on another (which there would be if two of those sounds were similar/the same), and that no set of samples or flight conditions has more influence on the fit than another. In total, there were 46 sUAS sounds, and 20 road vehicle sounds included in the regression.

A. Noise Metrics



(a) SEL_A and SEL_C (b) EPNL (c) L_5

Figure 6. Example noise metric calculations for a box truck drive–by, ID 418. The black and blue traces indicate the underlying time-varying metric calculations, and the red and green horizontal lines indicate the representative single values.

Noise metrics are signal processing techniques that are used to reduce a pressure-time history into a single number that is, perhaps, representative of the annoyance due to that sound. The set of metrics used here was arrived at after computation of the augmented regression model (described below) on a set of more than 10 metrics. Those shown here are meant to be both illustrative and to present the best results found thus far. Metric values were computed on the sample sounds recorded in the EER using the same equipment as the field recording, with the recording microphone placed near the center of the four subject locations, close to the SLM microphone used for calibration.

The four metrics shown are: A-weighted Sound Exposure Level (SEL_A), C-weighted Sound Exposure Level (SEL_C), Effective Perceived Noise Level (EPNL), and Loudness exceeded 5% of the time (L_5).[c] Examples of these metrics are shown in Figure 6 for sample ID 418 — a driveby of the box truck that has significant low–frequency energy. Briefly, these metrics are:

- SEL_A: This metric is the time integration of A-weighted sound energy, normalized to a duration of 1 second. The psophometric 'A' frequency weighting is the most commonly encountered measure of noise (and is implemented on nearly all hand-held sound level meters). dB_A is based on the pure-tone response of the human auditory system at a level of 40 Phon — corresponding, perhaps, to the level of a calm conversation in a very quiet setting. Despite not seeming applicable to loud/outdoor sounds, dB_A has found use as being correlated with annoyance across a large range of absolute levels. dB_A is typically averaged in time in order to smooth its response. The 'slow' averaging is used here, it is exponential with a time constant of 1 s. The computation of SEL_A is a straightforward Riemann sum of a dB_A over time.

- SEL_C: The computation of SEL_C is identical to SEL_A, except that the psophometric 'C' frequency weighting (dB_C) is used instead of dB_A. This corresponds to the human auditory system's frequency response at 100 Phon (corresponding more closely to a concert than a calm conversation), which is a significantly higher level than for A-weighting and more closely matches the levels at which the sample sounds were presented to the subjects. The result is that the SEL_C includes a significant amount of low- to mid-frequency energy (between, nominally, 100 and 1000 Hz), that SEL_A does not.

  This effect can be seen in the example in Fig. 6(a). Here, the sound generated by the vehicle is seen to have significantly more low–frequency energy included in SEL_C than in SEL_A.

- EPNL: This metric is used by regulatory bodies internationally to certify aircraft designs for community noise exposure.[22] This metric is a time-integration of a metric called the Tone-Corrected Perceived Noise Level (PNLT). PNLT is, in turn, based on one-third octave band data, sampled at half-second intervals. For a given interval, the one-third octave band data is converted into 'noy bands' by a non-linear transformation based on the response of the human ear at various absolute levels. A 'tone penalty' is computed between these bands — if adjacent bands have greatly differing magnitudes,

[c]For a more technical discussion of SEL_A, SEL_C, and EPNL, see Ruijgrok.[22] For information on L_5, see the DIN or ISO standard from which it is derived.[23]

there is a penalty associated with that difference. The noy bands are combined based on a empirical summation formula, and the tone penalty is added to create the PNLT value for that time interval.

The unit of PNLT is the PNdB. Due to the non-linearities in the PNLT computation, it is important to note that this a decibel-like quantity, and does not scale linearly with large changes in overall gain. For small changes in dB, the corresponding change in PNLT will be nearly equal, but large changes in dB can cause significant divergences between dB and PNdB.

EPNL is an integration of the PNLT time history for the period in which the noise is within -10 PNdB of the peak value. In this way, EPNL takes the duration of the noise into account. An example PNLT time history and EPNL value is shown in Fig. 6(b).

- L_5: There are several computational models that produce estimations of psychoacoustic loudness, a quantity akin to, but not directly related to, annoyance. In general, they are all based on models of the human auditory system. The current 'Zwicker' model is used in this study, as described by an international (ISO/DIN) standard.[23] By modeling the action of the various physical components of the ear, this model implicitly includes effects such as upward masking and non-linear frequency weighting, as well as some temporal effects. It does not include a penalty for tonality as EPNL attempts to do.[d] Its calculation results in a time-history, sampled at 100 Hz, in the decibel-like units of Phon. An example is shown in Fig. 6(c).

  Again, the metric must produce a single-number representation of a sound to be useful for linear regression. There is no widely-accepted single way to integrate loudness across time as there are for dB_A, dB_C, and PNLT. Instead, it is common practice to use quantiles of the loudness time history as a representative point.[24] In this case, it was determined that loudness exceeded 5% of the time produced results that were highly correlated with the mean subject response.[e] While this approach satisfies the single-value criterion, it means that L_5 will not account for duration effects.

In summation, SEL_A and SEL_C are the simplest measures, the difference between the two providing a glimpse of the importance of how various frequency components of the noise impact annoyance. EPNL is a more complex calculation that has widespread regulatory use, but does not reflect contemporary knowledge of the human auditory system. L_5 is the most accurate model of human hearing included, but may omit important aspects of the noise captured by the other metrics, such as corrections for tonality and duration. The performances of the various metrics on the data set may help to illuminate how the importance of these effects for understanding sUAS-noise annoyance.

B. Subject Responses

The regression analyses shown here were performed between the metric values of a sample sound and the arithmetic mean of the subject responses for that sound. It is important to observe how that mean relates to the spread of the subject responses for the samples. Figure 7 shows the histogram of the subject responses to the first sample sound (ID #1, an SUI flyover). It can be seen that the responses do not follow a simple bell-shaped distribution. Further, the responses encompass nearly the entire range of the scale. This indicates that there is a high-degree of variability between the subjects' impressions of the sounds.

There is exactly one response for each sample sound and for subject that participated in the test. For sample i and subject s, the response is $y_{i,s}$, where i is an index within the set of sample sounds found in Table 3 (and included in the regression analysis), and $s \in [1, 2, ..., S]$ for $S = 38$, the number of subjects that participated. For a given sample sound i, the arithmetic mean is:

$$\bar{y}_i = \frac{1}{S} \sum_{s=1}^{S} y_{i,s} \tag{2}$$

[d] There are other psychoacoustic measures based on calculations similar to loudness that attempt to account for this as well as other possible qualitative effects (e.g., roughness), however neither calculations of these measures, nor incorporation of these into a single explanatory model of human annoyance, are straightforward matters.[24]

[e] The 5% quantile is a commonly used value, though it is by no means an agreed-upon value across authors. In addition, the use of loudness in units of Sone — a power-like unit quantity of loudness — is also common for the quantile calculation.[15,24]

It is also important to estimate the confidence in the mean given the spread in the data. A confidence interval (CI) is a measure of the certainty in the estimate of the mean, and a measure of the power of the test to resolve this statistic. In general, if the confidence intervals of two samples overlap at all, it can be said that the responses are indistinguishably different given the power of the test to resolve the two (Analysis of Variance, a more rigorous test of this sort of question, is left out of this publication). The size of a CI is expected to shrink, roughly, as the square-root of the number of responses (subjects).

Although CIs typically require an assumption about the underlying distribution of the data — an assumption that would be dubiously made after observing Fig. 7 — CIs can still be constructed for the means of these samples via bootstrapping techniques. The method used to generate the intervals for Fig. 7, as well as all subsequent figures, was the Bias-Corrected, Accelerated method (BCa), as implemented in the MatLab bootci() function.[25] 100,000 bootstrap samples were used to compute each CI which, while perhaps excessive, ensures convergence of the bootstrapped interval values.

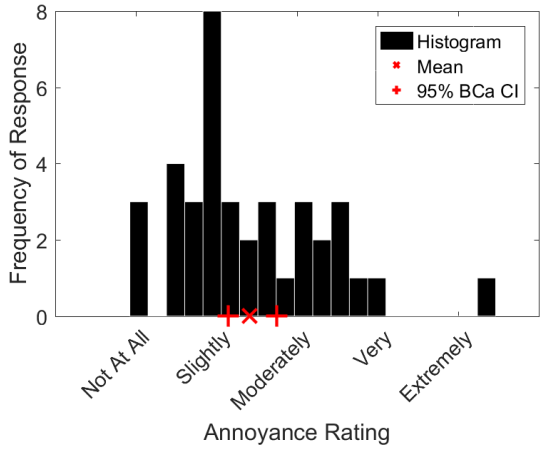


Figure 7. Histogram of subject responses for one of the baseline SUI samples (ID 1). Mean and BCa confidence interval are indicated.

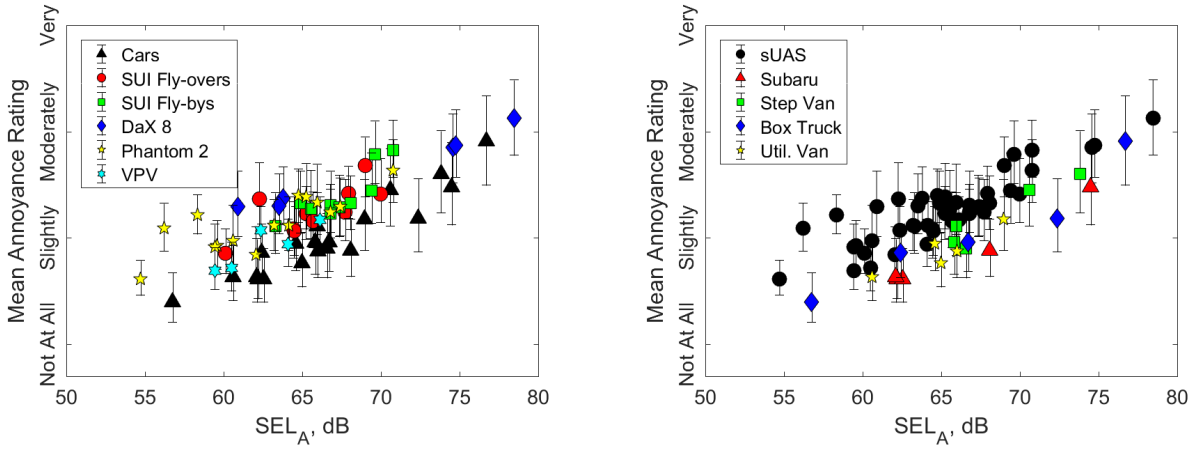


Figure 8. Scatter plots of mean subject responses and confidence intervals for the samples included in the regression analyses. The data in the two plots are the same, only the markers and colors are changed to aid the eye in vehicle identification.

Figure 8 shows the means and CIs for the samples included in the regression. The markers are colored to allow one to easily identify the corresponding vehicle. CIs generated via the BCa method are not constrained to be symmetric around the mean or the same size across samples. Both effects can be seen in this figure, and are due to the variation of skewness and central tendency in the subject data for the different samples respectively.

C. Pooled Regression

The first regression analyses were carried out by pooling all of the included samples into one set and fitting a line given the noise metric calculations on that set.[f] The model function is therefore:

$$\hat{Y} = \beta_0 + X\beta_1 \tag{3}$$

[f]Linear regression is a much written–about subject. Any basic book on statistics will cover the topic. The reader is directed to Chatterjee and Hadi[26] for a good introduction.

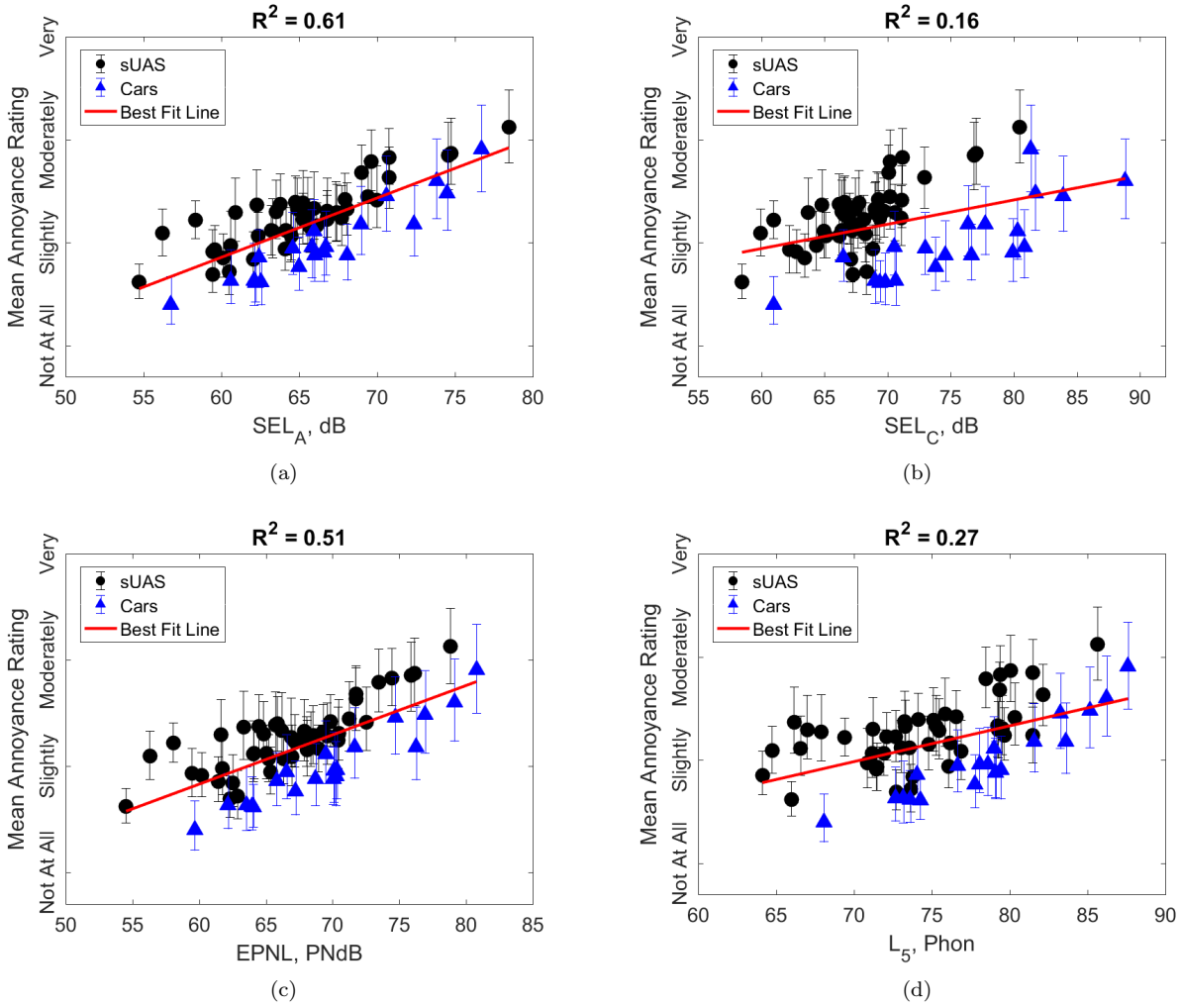


Figure 9. Pooled regression results for the four noise metrics.

where $\hat{Y}$ is a column vector of model-predicted mean responses to the sample sounds, X is the column vector of noise metric values for those samples, and β_0 and β_1 are the scalar regression coefficients — the results of the regression process. This process minimizes the sum of the squares of the difference between the model-predicted $\hat{Y}$ values and the column vector of the observed means $\bar{Y}$:

$$\min_{\beta_0,\beta_1} \quad \left(\bar{Y} - \hat{Y}\right)^T \left(\bar{Y} - \hat{Y}\right) \tag{4}$$

The MatLab fitlm() function was used to perform this minimization for all cases here. The results of these 'pooled' analyses are shown in Figure 9. Here, the samples corresponding to sUAS and road vehicles (Cars) are shown with different colors/markers, even though the line is fit to the totality of the data.

The main numerical result of these analyses is the square of the correlation coefficient R^2. This measures the percentage of the variance that was observed in $\bar{Y}$ that is accounted for by the model function, and is expressed as a number between 0 and 1. SEL_A is seen to offer the best performance. This is contrasted with SEL_C, which performs very poorly. This difference is primarily caused by the inclusion of the low–frequency energy from the road vehicle sounds in SEL_C, shifting the metric values for those sounds upward relative to the sUAS samples. This forces the regression line to flatten, causing the SEL_C model to account for a lower amount of the observed variance in the ordinate direction.

EPNL offers similar performance to SEL_A. The scatter of the individual data points is similar between Fig. 9(a) and (c), indicating that the two metrics are capturing similar aspects of the sounds. Previous studies have shown these two metrics to often be comparable for aircraft noise.[22]

L_5 offers surprisingly poor performance (at least for this first regression), though it can be seen that this is not caused by the same forces that result in poor performance for SEL_C. The sUAS and road vehicle sample sets, as measured in L_5, span similar loudness values, indicating that L_5 is simply failing to capture some important aspect of the noise.

Another important observation is that the road vehicles seem to be systematically judged to be less annoying than the sUAS for the same amount of metric noise. This seems to hold true for all of the metrics. This observation will lead the development of the next regression model.

D. Regression Diagnostics

Having made these initial observations, it is important to address the underlying assumptions of the linear regression method, and to make sure that they are satisfied for this data set. Beyond the expectation that the data would follow a linear trend given the form of the question, there are three main criteria for applicability of linear regression. These are: independence of the samples, normal-distribution of the residuals (with 0-mean), and homoskedasticity — that the distribution of the residuals does not vary systematically with the predictors.[26]

Regarding independence, this was provided, insofar as it can be provided in human subject testing, by the test protocol and design: All subjects went through familiarization and training sessions before the test began in order to acquaint the subjects with the range of the sounds and the use of the response scale. Further, the randomization of the test order for each group of subjects provided that the effects of learning/fatigue would be well–distributed among the groups. Lastly, all subjects responded to all questions, and, for these analyses, no multiple responses were grouped into a single mean.

The last two requirements are typically fulfilled by observation of plots of the residuals of a regression. A histogram of the residuals can provide confidence that they are following a bell-shaped distribution (the requirement of normality being relatively weak[17]). Heteroskedasticity, as well as other unwanted systematic effects in the data, can be seen by plotting the residuals against their underlying metric values. In all cases, these plots were generated and showed that the assumptions of the regression method were sufficiently met.

E. Augmented Regression

Given the observed systematic differences between the mean annoyance values of the sUAS samples and the road vehicle samples, an augmented linear model was proposed. A binary term C was added to the model equation:

$$\hat{Y} = \beta_0 + X\beta_1 + C\beta_2 \tag{5}$$

c_i, the element of C corresponding to sample index i, is 1 if i corresponds to a road vehicle sample and 0 if it corresponds to an sUAS sample. This effectively allows for two lines to be fit to the data: one to the sUAS data, and one to the road vehicle data. These lines are constrained to have the same slope, as there is not a compelling reason that, asymptotically, they should be different. Thus a single slope is determined for all of the data while β_2 captures the offset between the two lines in the ordinate direction. This offset, as measured in the units of the metric, is given by β_2/β_1. A sample result of regressing this model on the SEL_A data is shown in Figure 10. The results of using this model on all 4 metrics are given in Table 4.

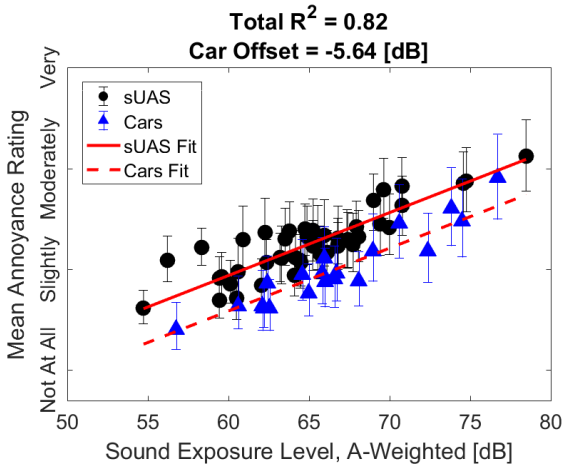


Figure 10. Regression results for the augmented linear model for SEL_A.

The explanatory value of the model for all metrics, in terms of R^2, is greatly improved. The value of the offset is not a small number for any of the metrics, indicating that there is a significant amount of subjective difference between the sUAS and road vehicle sounds that is not captured by any of the metrics. The inclusion of this binary predictor in the regression is shown to be very significant in all cases — the p-values of the t-tests for inclusion of this predictor are well below the canonical 95% confidence value of .05 in all cases.

Table 4. Regression results for the augmented linear model. Offset is measured in the respective metric's unit.

Metric	R^2	Offset
SEL_A	.82	5.6
SEL_C	.68	12.8
EPNL	.80	7.6
L_5	.75	7.5

F. Bootstrapped Regression

The last form of regression analysis deals with the determination of confidence intervals for the R^2 and offset values that have been determined for the augmented model. These interval estimates will help to indicate whether, given this dataset, one metric can confidently be resolved from another, and whether the offset value is significantly different from 0.[g]

To generate these intervals, a non-parametric bootstrapping technique was used, similar to the BCa approach used to bootstrap CIs to the means. For each bootstrap sample b, the original mean subject response vector was modified by resampling, with replacement, from the underlying subject data, and calculating new means to form a resampled $\bar{y}_b$. The new vector $\bar{Y}_b$ was then used to perform a linear regression (as in Eq. 4). The result of each bootstrap sample is a $\beta_{0,b}$, $\beta_{1,b}$, and $\beta_{2,b}$ that describe the best fit augmented model to the resampled means. This process is repeated many times to form the non-parametric distribution of these βs. This technique is a form of the 'percentile' bootstrapping method, and is similar BCa, though slower to converge.[27h]

Once the sets of bootstrap βs have been assembled, the 2.5% and 97.5% percentiles are taken of the sets, and presented as the two-sided 95% confidence intervals. 20,000 bootstrap samples were drawn for the results shown here, which produced convergence of the CIs to one part in 1,000 in the interval estimates — a process that took about 30 minutes per metric on one core of a contemporary laptop. The results are shown in Table 5, and graphically in Figure 11.

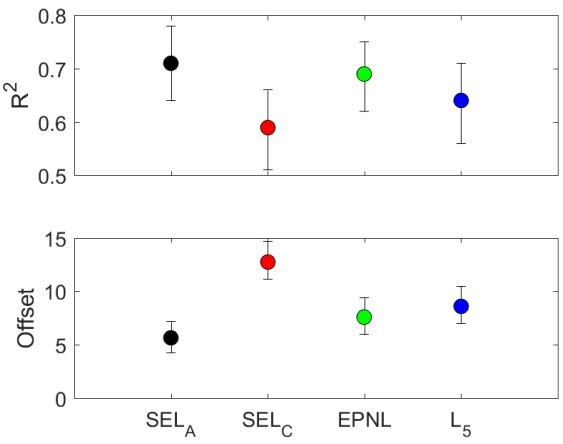


Figure 11. Bootstrapped 95% confidence intervals for the augmented linear regression model. Offset is measured in the respective metric's unit.

First, it is important to observe that the mean values for R^2 are reduced by bootstrapping. This is due to the fact that this method effectively reintroduces the variance behind the means of the original $\bar{Y}$ vector, so that there is more variance, but the same best fitting model. The intervals for R^2 preserve the same trends as before: integrated metrics tend to perform better, and the extra low-frequency energy in SEL_C causes it to perform poorly. The overlap in confidence intervals in the top half of Fig. 11 indicates this dataset cannot be used to confidently determine the best of the four metrics (though such a claim based on the results of a single subjective test would be dubious regardless of this outcome).

In contrast, the fact that none of the confidence intervals in the lower half of Fig. 11 contain 0, is an indication that none of the four metrics sufficiently captured some subjective aspect of sUAS relative to road vehicle noise. The apparent negative covariance of R^2 and offset seen in Fig. 11 indicates a truism: the better a metric performs, the smaller the offset between sUAS and road vehicles.

[g]The latter result should be expected given the result of the t–tests of the β_2s, but there is still a great difference between the determination of a 5 dB offset ± 4 dB, and 5 ± 1 dB.

[h]There is a parametric method, implemented in the MatLab regression package, to generate confidence bounds on regression coefficients using the covariance matrix thereof. Although this method produces similar, though unequal, intervals for the offsets, it does not produce estimates for R^2. Further, this parametric method does not address the underlying distributions behind the $\bar{y}_i$ means.

Table 5. Bootstrapped 95% confidence interval results for the augmented linear regression model. Offset is measured in the respective metric's unit.

Metric	Median R^2	R^2 CI	Median Offset	Offset CI
SEL_A	.71	[.64, .78]	5.64	[4.3, 7.2]
SEL_C	.59	[.51, .66]	12.75	[11.1, 14.6]
EPNL	.69	[.62, .75]	7.58	[6.0, 9.4]
L_5	.64	[.57, .72]	8.60	[5.9, 9.3]

Discussion of Regression Results

The idea that various sources of noise may elicit significantly different annoyance responses is not unprecedented. A germane example, though formulated for multiple-event annoyance, is work by Miedema and Oudshoorn that showed statistically significant differences in the DNL (an aggregate noise metric) response between road noise, rail noise, and aircraft noise.[28]

In the present study, the subjects responded to single events of noise, and were not instructed on what the sources of the noise may have been. Common comments during informal conversations after the test were that the subjects were typically not aware that the non-road noises came from 'drones,' and that the fact that some were flying overhead and some were presented as drive-bys did not significantly impact their judgments. This suggests that the subjects were queuing off of qualitative differences between the sample sounds.

If the difference between the effects of sUAS noise and road vehicle noise are qualitative, then the primary implication is that the use of contemporary noise metrics for the evaluation of sUAS noise may have to include a significant component for the qualitative aspects of sUAS noise *vis a vis* the noises for which those metrics are commonly used. Those who expect for sUAS to gain widespread community acceptance based on the idea that they will produce no more annoyance than the equivalent amount of traffic noise, may not be correct.

An important caveat is that, as stated in the introduction, this test was not conceived to be a comprehensive examination of noise from either sUAS or road vehicles. Rather, it was meant, primarily, to demonstrate the extensibility of tools and facilities that NASA already possesses to the realm of sUAS noise. Therefore, it is unwise to attempt to generalize the results of this study beyond those stated in the discussion, and beyond the limited set of vehicles and conditions tested.

G. Regarding Height

Finally, an interesting effect has been noticed in the data, though not completely objectively explored at the time of writing. For sUAS samples between which only height varies (where the other parameters of the operations are held constant) there is usually insignificant change in the annoyance response. An example of this can be seen in Figure 12. The figure shows the annoyance responses for the 2-bladed SUI vehicle flown at four different altitudes, from 20 to 100 m AGL. There is no significant difference between the annoyance responses (the confidence intervals all overlap), even though there is roughly an 8 dB difference in SEL_A between the samples. Similar trends exist for the other sets of sUAV sounds in which parameters other than height are held nominally constant.

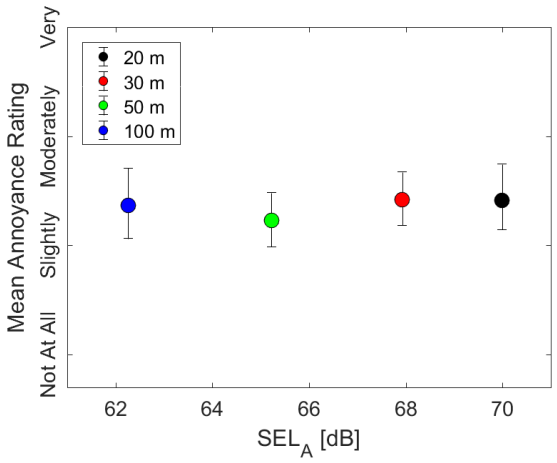


Figure 12. The effect of changing height on annoyance. The four samples shown are all from the SUI Endurance equipped with 2–blade props, flying at 5 m/s, and at different heights AGL (IDs 1, 5, 17, and 20).

There are two important points to be made regarding this observation:

1. Alleviating annoyance from sUAS operations may not be simply a matter of flying higher. The maximum dB_A of a flyover would be expected to reduce by 6 dB every time the distance between the source and receiver doubles. SEL_A, as a time-integrated metric, would be expected to reduce less than this per doubling of distance, as can be seen in Fig. 12.

   A common comment from subjects was that sounds which appeared to 'loiter' were judged more harshly than those that didn't.

2. This result could, at least partially, work toward explaining the offset between sUAS and road vehicle annoyance observed in the regression results. The road vehicles were all recorded at (nominally) 10 m distance, and moving at and 10 m/s. This means that while the closest/fastest sUAS recordings were geometrically comparable to all of the road vehicles, the vast majority of sUAS recordings were farther/slower. The results in Fig. 12 suggest that the addition of a loitering penalty, formulated only from geometrical considerations, and within the limits of this dataset, will work to alleviate some, if not all, of the observed difference between sUAS and the road vehicles.

   Given that the result of Fig. 12 seems to indicate that this penalty ought to be able to explain a difference in annoyance corresponding to 8 dB SEL_A, and the offset as measured in SEL_A is confident to only 4.3 dB, it is not unreasonable to assume that the proper formulation of such a loitering correction could account for the entirety of the significant region of the offset.

The result of equal annoyance with distance should not be construed as being extensible beyond the bounds of this test. It is highly likely that, given that the car sounds were the fastest/closest, and the subjects were known to (at least in part) develop cognitive heuristics to judge sounds, any heuristic penalty associated with duration would be anchored by the car sounds and then extended as a penalty to those that are slower/farther (all of the sUAS sounds). Likewise, it is possible, given that cars driving by are a common occurrence in many people's daily life (especially in Hampton Roads, VA, where all of the test subjects were drawn from), and car ownership is necessary for many there as well, that proximate road noise constitutes a sort of cognitive baseline for acceptable noise. Lastly, on the opposite end of the spectrum, it is known that there is a startle–related onset penalty for noises that rise in intensity too rapidly, so that the concept of a time–based correction to noise beyond that provided by a simple time integration of the signal, is not unprecedented.[29]

This possibility will be the subject of further data analysis and research. Explaining the offset in terms of a loitering penalty would not exonerate sUAS noise from the previous discussion; rather, the implication is that for sUAS operators to compete on a level playing field noise-wise, they will have to give a good deal of credence to the fact that the speeding up of their operations is going to be the key to making their noise acceptable. This is opposed to giving credence to the idea that there is a qualitative component of sUAS noise that is offensive and not (as) present in road vehicle noise.

H. Future Analyses

The path forward for this work is clear in the near term. Analysis of the current data set will continue along 3 fronts, related to the three research questions outlined previously:

1. Analysis of variance on the samples that were presented in repetition to the subjects. These include the SUI repeats — both when the identical sounds were repeated, as well as identical conditions — and the sUAS auralization samples. This will primarily shed light on the inter- and intra-subject components of the variance in the dataset and help to inform the design of future such tests.

2. Factor analyses using ANOVA. This can be used to explore relationships (or the lack thereof) between the 'factors' of the samples (e.g., speed and distance) and the resultant annoyance. This work will inform component-level design considerations for sUAS in regards to noise.

3. Work to increase the explanatory power of the noise metrics employed for this effort. This will encompass any effort to produce a loitering correction as described above, but can also be used to bring further psychoacoustic measures (such as tonality) to bear. This work would go toward creating tools that can predict annoyance due to a wide range of sUAS operations noise.

V. Conclusions

This paper describes the results of a recent psychoacoustic test at NASA Langley to explore differences in subjective response to noise from flyovers of small unmanned aerial systems (sUAS) with noise from drive-bys of road vehicles encountered in residential neighborhoods. Recordings of the various vehicles were collected during the second half of 2016 and early 2017. The recordings were used, along with auralizations previously created for other tests/purposes, in a psychoacoustic test in February 2017. This test took place in the Exterior Effects Room, a calibrated 3D sound environment and human subject test facility. Subjects provided holistic responses regarding their levels of annoyance to the various test sounds. Data from 38 subjects for all test sounds were collected.

Initial analysis of the data from this test indicates that there may be a systematic difference between the annoyance response generated by the noise of the sUAS and the road vehicles included in this study. It is unknown as of now whether this difference can be accounted for by other factors, or whether it is being generated by qualitative differences between the sound of road vehicles and sUAS. This result casts doubt on the idea that sUAS operators can expect their operations to be greeted with minimal noise-based opposition as long as the sound of their systems are "no louder than" conventional package delivery solutions.

Further analysis of the data is ongoing, including factor analysis and interrogation of the penalty from the point of view that it is resulting from a loitering sensation produced by the sUAS operations. A follow-on test, informed by the results of this test, is being planned for later in 2017.

Acknowledgments

A great deal of credit is due to the many people who aided this effort over the past year — despite the fact that only two appear as authors of this report.

Matt Hayes, a technician in the Structural Acoustics Branch at NASA Langley, was immensely helpful throughout all of the recording efforts. He was the only person to be present at all of the recordings, squiring the data collection system around the country. Other people who were instrumental for the recordings included those who helped with the data collection system, the NASA and SUI personnel who operated the sUAS, and the Alutiiq personnel who drove the road vehicles. The DEP auralizations were provided by Dan Palumbo.

A similarly large number of people were involved in human subject testing portion of this effort. Among those involved in pilot testing, debugging, and actually operating the EER for subjects were: Menachem Rafaelof, Aric Aumann, Kevin Shepherd, and Brian Tuttle (who deserves credit for the 'acoustic fisheye'). Regina Johns and Erin Thomas were the contractors in charge of supplying the test subjects.

Lastly, Colin Theodore, director of the DELIVER project, deserves a good deal of credit for, at the very least, putting up with the vicissitudes of this projects' timeline.

References

[1] "Fact Sheet - Small Unmanned Aircraft Regulations (Part 107)," https://www.faa.gov/news/fact_sheets/news_story.cfm?newsId=20516, Accessed: April 2017.

[2] "Amazon Prime Air (Commercial)," https://www.youtube.com/watch?v=MXo_d6tNWuY, Accessed: April 2017.

[3] Shepherd, K. P., "The Subjective Evaluation of Noise from Light Aircraft," NASA CR-2773, Prepared by The University of Utah for Langley Research Center, Salt Lake City, UT, 1976.

[4] "Drone America - DAx8," http://www.droneamerica.com/systems/dax8, Accessed: April 2017.

[5] "Stingray 500 - CJ Youngblood Ent." http://www.curtisyoungblood.com/legacy-product-support-curtis-youngblood/attachment/stingray-500/, Accessed: April 2017.

[6] "DJI Phantom 2," http://www.dji.com/phantom-2, Accessed: April 2017.

[7] "3DR Pixhawk," https://store.3dr.com/t/pixhawk/, Accessed: April 2017.

[8] Department of Defense, "Global Positioning System Standard Positioning Service Performance Standard," Tech. rep., DoD Positioning, Navigation, and Timing Executive Committee, September 2008.

[9] Zawodny, N. S., Jr, D. D. B., and Burley, C. L., "Acoustic Characterization and Prediction of Representative, Small-Scale Rotary-Wing Unmanned Aircraft System Components," *Proceedings of the 72nd American Helicopter Society Forum*, AHS, West Palm Beach, FL, 2016.

[10] Cabell, R., Grosveld, F., and McSwain, R., "Measured noise from small unmanned aerial vehicles," *Proceedings of NOISE-CON 2016*, Vol. 252, Institute of Noise Control Engineering, Providence, RI, 2016, p. 345354.

[11] "EVK-7 — u-blox," https://www.u-blox.com/en/product/evk-7, Accessed: April 2017.

[12] "Leap Seconds," http://tycho.usno.navy.mil/leapsec.html, Accessed: April 2017.

[13] "Tascam DR-701D," https://tascam.com/product/dr-701d/, Accessed: April 2017.

[14] Christian, A. and Lawrence, J., "Initial Development of a Quadcopter Simulation Environment for Auralization," *Proceedings of the 72nd American Helicopter Society Forum*, Vol. 1, AHS, West Palm Beach, FL, 2016, pp. 57–67.

[15] Rizzi, S. A. et al., "Perceived annoyance to noise produced by a distributed electric propulsion high lift system," *Proceedings of the 2017 AIAA Aviation and Aeronautics Forum and Exposition*, The American Institute of Aeronautics and Astronautics, Denver, CO, 2017.

[16] Rafaelof, M., "A Model to Gauge the Annoyance of Arbitrary Time-varying Sound," *Proceedings of NOISE-CON 2016*, Institute of Noise Control Engineering, Providence, RI, 2016.

[17] Bech, S. and Zacharov, N., *Perceptual Audio Evaluation*, John Wiley & Sons, Hoboken, NJ, 2006.

[18] Montgomery, D. C., *Design and Analysis of Experiments*, John Wiley & Sons, New York, NY, 3rd ed., 1991.

[19] Faller II, K. J., Rizzi, S. A., and Aumann, A. R., "Acoustic performance of a real-time three-dimensional sound-reproduction system," TM -2013-218004, NASA, Hampton, VA, 2013.

[20] ISO/TC 43 (Acoustics), "ISO 389-1: Reference zero for the calibration of audiometric equipment: Reference equivalent threshold sound pressure levels for pure tones and supra-aural earphones," International Organization for Standardization, 1998.

[21] Fields, J. M. et al., "Standardized General–Purpose Noise Reaction Questions for Community Noise Surveys: Research and a Recommendation," *Journal of Sound and Vibration*, Vol. 242, No. 4, 2001, pp. 641–679.

[22] Ruijrok, G. J. J., *Elements of aviation acoustics*, VSSD, Delft, The Netherlands, 2nd ed., 2007.

[23] ISO/TC 43 (Acoustics), "ISO 532-1: Methods for calculating loudness: Zwicker method," International Organization for Standardization, 2015.

[24] More, S. R., *Aircraft Noise Characteristics and Metrics*, Ph.D. thesis, Purdue University, West Lafayette, IN, July 2011.

[25] Efron, B., "Better Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, Vol. 82, 1987, pp. 171–200.

[26] Chatterjee, S. and Hadi, A. S., *Regression Analysis by Example*, Wiley, Hoboken, NJ, 5th ed., 2012.

[27] Carpenter, J. and Bithell, J., "Bootstrap confidence intervals: when, which, what?" *Statistics in Medicine*, Vol. 19, 2000, pp. 1141–1164.

[28] Miedema, H. M. E. and Oudshoorn, C. G. M., "Annoyance from Transportation Noise: Relationships with Exposure Metrics DNL and DENL and Their Confidence Intervals," *Environmental Health Perspectives*, Vol. 109, No. 4, 2001, pp. 409–416.

[29] Plotkin, K. J. et al., "The Effect of Onset Rate on Aircraft Noise Annoyance, Vol. 1: Laboratory Experiments," WR 91-19, Wyle Research, Arlington, VA, 1991.